Point-triplet descriptors for 3D facial landmark localisation

Marcelo Romero[1], Nick Pears[2], Adriana Vilchis[1], Juan C. Ávila[1], Otniel Portillo[1]
Autonomous University of the State of Mexico [1]
{mromeroh, avilchisg, jcavilav, oportillor}@uaemex.mx
The University of York, UK [2]
nep@cs.york.ac.uk

Abstract

An investigation to localise facial landmarks from 3D images is presented, without using any assumption concerning facial pose. This paper introduces new surface descriptors, which are derived from either unstructured face data, or a radial basis function (RBF) model of the facial surface. Two new variants of feature descriptors are described, generally named as point–triplet descriptors because they require three vertices to be computed. The first is related to the classical depth map feature, which is referred to as weighted–interpolated depth map. The second variant of descriptors are derived from an implicit RBF model, they are referred to as surface RBF signature (SRS) features. Both variants of descriptors are able to encode surface information within a triangular region defined by a point–triplet into a surface signature, which could be useful not only for 3D face processing but also within a number of graph based retrieval applications. These descriptors are embedded into a system designed to localise the nose–tip and two inner–eye corners. Landmark localisation performance is reported by computing errors of estimated landmark locations against our respective ground–truth data from the Face Recognition Grand Challenge (FRGC) database.

1. Introduction

The human face is a huge source of information, and it plays an essential role in social interactions. Physically speaking, the face is a natural human way of identification, conveying race, age and gender; and for the people who frequently interact with each person (such as colleagues, friends, and family), the person's face is closely associated with all that he/she is. Behaviourally speaking, the face is a primary actor within interpersonal communication, essentially, because it is the means of expressing emotions [6]. Furthermore, it has been said that the effectiveness when a message is transmitted is 7% from spoken words, 38% from voice intonation, and 55% from facial expressions, which implies that facial expressions are the main modality in human communications [7]. These are some facts that motivate the research community to study the human face from different perspectives.

For many face processing algorithms: face animation, face registration, face alignment, face recognition and verification; accurate facial landmark localisation is an essential precursor. For instance, it is well known that even holistic matching methods, such as *Eigenfaces* [8] and *Fisherfaces* [9], need accurate locations of key facial features for face pose normalisation; where noticeable degradation in recognition performance is observed without accurate facial feature locations. Furthermore, it is generally believed that, an improved landmark localisation will increase the effectiveness of many face processing applications [11-15].

After several years of research, face processing has become an everyday tool in real life applications [11]-[15]. However, convincing solutions for 3D data that work well over a wide range of head poses are still needed. This paper presents some progress in localising facial landmarks within 3D face data, without any assumptions concerning facial pose.

The rest of this paper is structured as follows. Section 2 introduces the *point-triplet descriptors*. Section 3 presents the experimental framework to localise facial landmarks. Section 4 shows the experimental framework and results. Finally, Section 5 discusses and concludes this paper.

2. Point-triplet feature descriptors

This section describes the *point–triplet feature descriptors*, which given a triplet of 3D points, are able to encode a 3D shape contained in the triangular region defined by this triplet into a surface signature. It presents two variants of point–triplet descriptors. The first is related to the classical depth map feature, this feature is referred to as *weighted–interpolated depth* *map*. The second variant of descriptors are derived from an implicit *radial basis function* (RBF) model, they are referred to as *surface RBF signature* (SRS) features, which are related to the previous work in sampling an RBF model [1]. Both variants of descriptors are a natural extension of the previous work in landmark localisation [1]-[5]. The point-triplet descriptors are able to encode surface information within a triangular region defined by a point–triplet into a surface signature, which could be useful not only for 3D face processing but, also, within a number of graph based retrieval applications.

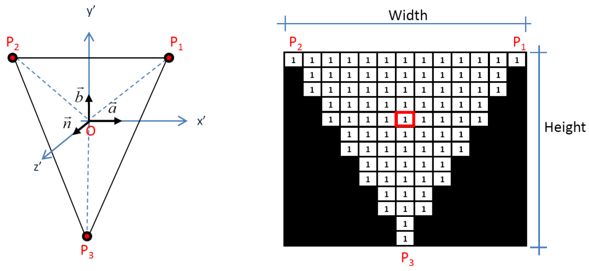
However, this paper evaluates their ability to identify point–triplets of facial landmarks, two endocanthions (inner-eye corners) and a pronasale (nose-tip), as a first application. To do this, first generate candidate landmark–triplets as follows: for every vertex, *DLP* and *SSR value features* [4] were computed, and only those within three standard deviations were retained. Then, using contextual support [2], a pair of candidate landmarks is created. As long as SSR value features robustly detect the pronasale landmark, it was found that many candidate pairs of endocanthions can be deleted, as no pronasale landmarks support them [5]. After this, only candidate landmarks with the minimum Mahalanobis distance to the mean of training SSR value features, within a radius of 10 mm, are kept. This is found necessary to reduce the potential number of candidate triplets. Unique combinations of endocanthions and pronasale landmarks, with mutual contextual support, were then created, using a right-hand orientation, from the left to the right endocanthion, and then to the pronasale landmark. Such orientation was defined using the normal to the plane defined by each triplet, which was oriented towards the camera's viewpoint. At the end of this process, a practical number of candidate point–triplets for every testing face were obtained, to which, the *point–triplet descriptors* were applied. The following subsections define the *point–triplet descriptors* and the experimental evaluation used to identify triplets of facial landmarks.

2.1 Weighted–interpolated depth map

A weighted–interpolated depth map is a point–triplet descriptor closely related to the classical *depth map feature*. The idea here is to compute a depth map using a point–triplet which effectively defines a triangular–region within a surface as illustrated in Figure 1(a). Given a triplet of 3D points (p_1, p_2, p_3), a weighted–interpolated depth map is computed as follows. Firstly, the baricenter of the triangular–plane is computed, and this point is used as the origin O. From this origin and based on the plane's normal ($\bar{n}$ in Figure 1-a) a local right-hand basis is defined, which is oriented towards the camera's viewpoint. Then, a [13x13] regular grid is created, but only those points within the triangular region are used. To do this, a binary mask is used, as shown in Figure 1(b). Then, for each sampling point within this triangular mask, a depth is estimated by using *inverse square weighted interpolation* $f(x, y)$:

$$f(x,y) = \frac{\sum_{i=1}^{n} \frac{f(x_i, y_i)}{R_i^2}}{\sum_{i=1}^{n} \frac{1}{R_i^2}} \tag{1}$$

where $f(x_i, y_i)$ is the depth value in the (x_i, y_i) coordinate, and $R_i^2 = (x - x_i)^2 + (y - y_i)^2$. To do this, neighbouring points in a radius $r = \sqrt{dw^2 + dh^2}$ are collected, where $dw = width/12$ and $dh = height/12$. In this definition, *width* and *height* are the Euclidean distance from p_1 to p_2, and from the middle–point of (p_1, p_2) to p_3, respectively.



(a) (b)

Figure 1. A weighted–interpolated depth map is computed by generating a [13x13] regular grid, then applying a binary mask.

2.2 Surface RBF Signature (SRS) Features

In this subsection four alternate features to analyse a 3D shape given a triplet of 3D points are presented. All of them use a radial basis function (RBF) model to compute depths. Thus, this family is referred to as surface RBF signature (SRS) features, namely: baricenter depth map, 7–bins SRS vector, SRS depth map, and SRS histogram.

The idea here is to sample an RBF model by a set of n points which lie within the triangular–region defined by (p_1, p_2, p_3). There are several ways to generate such sets of sampling points, beginning with the classical approach to computing a depth map using a regular grid. However, the point of interest is the shape enclosed by this triplet of points. Hence, only points within this triangle are considered here, in the first approach, this is done by using a binary mask, see Figure 1(b).

However, a triplet of non–colinear points which define a triangle is expected. Taking advantage of their geometry, it is then straightforward to compute their baricenter O (see Figure 2). Furthermore, it is clear that this process can be done iteratively. This procedure is referred to as a *baricenter sampling points algorithm* [5] which motivates the computation of the SRS descriptors: baricenter depth map, 7–bins SRS vector, and SRS histogram, introduced in the following subsections.

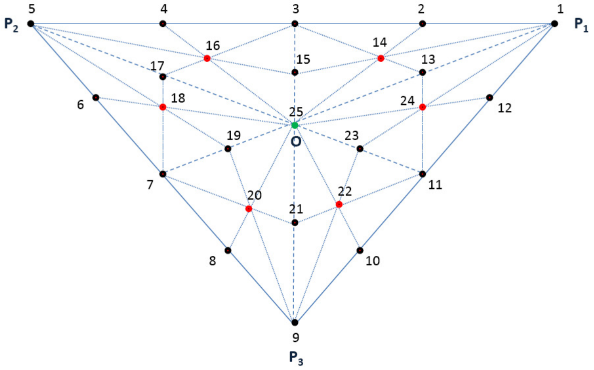


Figure 2. Labelled sampling points, computing baricenters from a triangular region in 2 iterations [5].

a) Baricenter Depth Map

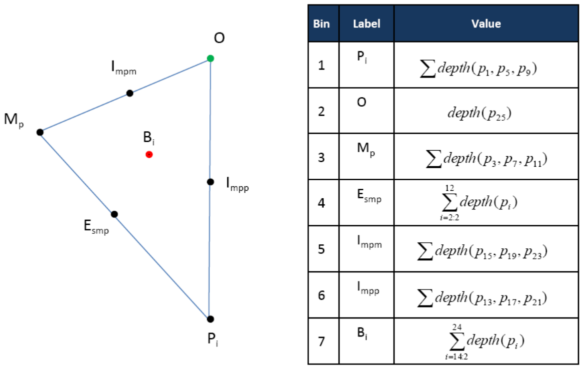
A baricenter depth map is a straightforward solution, which is generated from sampling points using the baricenter–based algorithm with two iterations. Figure 2 shows 25 labelled sampling points generated using the baricenter approach with 2 iterations. It is known that these sampling points will be the same no matter how the three points within the triplet are sorted. However, in order to encode depths from these sampling points they are labelled as shown in Figure 2. Then, the labels are used to assign each depth (Distance to Surface, DTS, value [1]) into a specific bin as indicated in Table I. As observed, this is a pose–invariant solution, but it is oriented, and different features are obtained if the triplet (p_1, p_2, p_3) is sorted differently, which affects labels in Figure 2.

b) 7–bins SRS Vector

A *7–bins SRS vector* is a feature descriptor which, contrarily to a *depth map*, is pose–invariant and undirected, which make this an attractive descriptor for several applications. Such a feature vector is a straightforward solution computed from 25 sampling points, as detailed in Figure 2, generated from the baricenter algorithm in 2 iterations. The idea here is to fold down the initial triangular section, collapsing symmetrical points into just one, for example: points p_1, p_2 and p_3; and the internal baricenters. This descriptor is inspired by an ideal model, an equilateral triangle that can be folded down symmetrically. In this case, it is done by adding depths of what were considered coincident points in the ideal model. Addition is considered an appropriate operation because it is commutative, making an undirected feature descriptor. Figure 2 illustrates the 25 sampling points, where labels in this case are just for reference to show how they are folded down, into a new triangular region as observed in Figure 3. Using this approach, depths in Figure 3 are distance to surface values (DTS) from each sample point to the surface RBF model.

TABLE I. BINNING DTS VALUES FOR BARICENTER DEPTH MAPS.

5	4	3	2	1
17	16	15	14	13
6	18	25	24	12
7	19	21	23	11
8	20	9	22	10



Bin	Label	Value
1	P_i	$\sum depth(p_1, p_5, p_9)$
2	O	$depth(p_{25})$
3	M_p	$\sum depth(p_3, p_7, p_{11})$
4	E_{smp}	$\sum_{i=2:2}^{12} depth(p_i)$
5	I_{mpm}	$\sum depth(p_{15}, p_{19}, p_{23})$
6	I_{mpp}	$\sum depth(p_{13}, p_{17}, p_{21})$
7	B_i	$\sum_{i=14:2}^{24} depth(p_i)$

Figure 3. The 7–bins SRS vector is generated by folding down 25 sampling points from the baricenter algorithm (2 iterations), where *depth*$_i$ is the distance to surface value from the *i*–sample point to the surface's RBF model.

c) SRS Depth Map

An *SRS depth map* is a counterpart to the *weighted–interpolated depth map* (Section 2.1), where depths are generated by sampling an RBF model using a regular grid, but taking only those values within the triangular region defined by a point–triplet (p_1, p_2, p_3). This is possible by applying a type of binary mask (see Figure 1), however, this solution is neither undirected nor pose–invariant.

d) SRS Histograms

A *surface RBF signature (SRS) histogram* is related to Pears' *SSR histograms* [1]. Given a point–triplet which defines a triangular region in the 3D space, an *SRS histogram* is computed by generating sampling points using our baricenter algorithm. Distance to surface (DTS) values are then obtained from this sample set at different heights, above and below the target triangular region. Normalised DTS values are obtained by dividing each DTS by its respective height, producing values between *-1* to *1*. Finally, a 23–bin histogram is produced with the normalised DTS values for each height. In doing this, consistent triangular regions from views at different heights are being sought [5]. The theory being that given a triangular region defined by a point–triplet, an SRS histogram is computed by sampling an RBF surface model at different heights, where such a sampling set is produced using our *baricenter sampling point algorithm* mentioned in Section 2.2.

3. Landmark Localisation

This section presents the experimental framework to illustrate how the *point–triplet descriptors* can be used to identify distinctive facial landmarks, the pronasale and endocanthions. As shown in Figure 4, our investigation firstly needs candidate point–triplets. To do this, *distance to local plane (DLP)* and *spherically sampled RBF (SSR) value* features are used [3]-[5], along with contextual support based on Euclidean lengths [2]. *Point–triplet descriptors* are then computed and the candidate triplet with the minimum Mahalanobis distance to the mean of respective point–triplet training data is stored for localisation performance evaluation.

3.1 Testing Procedure

As observed in Table II, a system was created using each point–triplet descriptor to localise the pronasale and endocanthion landmarks, giving five point–triplet systems (PT–S) in total, which are tested as illustrated in Figure 4. The experimental procedure is as follows:

1. Separate training and testing sets from the Face Recognition Grand Challenge (FRGC) database [10] are defined. Particularly, only data with 2D-3D correspondence [5] from Spring-2003 subset, which present variations in depth but generally neutral expressions. Thus, 200 shape images from different people for training and 509 faces from different people for testing are used.
2. From these 200 training images, point–triplet training data is gathered at the ground–truth [2] level.
3. For each testing face above, candidate triplet–landmarks (endocanthions and pronasale) are collected as illustrated in Figure 4. Firstly, initial candidate lists for endocanthions and pronasale landmarks are collected. This is done by computing *distance to local plane* (DLP) first, and then, *spherically sampled RBF* (SSR) values for every vertex within a testing face. For a vertex to be a candidate, it must be within 3–standard deviations of respective training data. Secondly, point–pair candidates were gathered based on training Euclidean distance within three standard deviations. This produces both candidate endocanthion–pairs and endocanthion–pronasale–pairs. This allows endocanthion pairs (left–right) without pronasale support to be ignored, as they are not useful for creating triplets. Candidates with the minimum Mahalanobis distance to the mean of SSR value training data are then kept, giving a kind of local maximum and local minimum for pronasale and endocanthion landmarks. Finally, a triplet is formed by combining pronasale and endocanthion candidates mutually supported.
4. Every triplet is right-hand oriented, from left to right endocanthion, then to the pronasale candidate. This allows identification of duplicated triplets, which are expected from the shape similarity between the left and right endocanthions.
5. Depths for *weighted–interpolated depth maps* are computed using raw points within the triangular region defined by the candidate triplet (p_1, p_2, p_3).
6. *SRS depth maps* are produced by computing 25 sampling points, using the *baricenter algorithm* with 2 iterations, then binning each depth into a [5x5] array, as illustrated in Table I.
7. *7–bins SRS vector* features are computed as defined in Section 2.2(b).
8. *SRS histogram* features are generated from sample points computed by 4 iterations of our *baricenter algorithm*, 8 heights: 10:5:45 and 23 bins, giving *SRS histograms* of [23x8].
9. When appropriate, *Principal Component Analysis* (PCA) is used to reduce the feature space to 8, 16, 32, and/or 64 dimensions.
10. Point–triplet features are computed for every candidate triplet and compared against respective training data. Then, the triplet with the minimum Mahalanobis distance to the mean of respective

point–triplet training data is taken as the best landmark estimation.

11. Localisation performance figures are generated by computing localisation errors between estimated landmarks against our manually marked ground–truth [2] from the FRGC database. The results are then used to present localisation performance figures, using thresholds in Table III for successful, poor, and failure localisations.

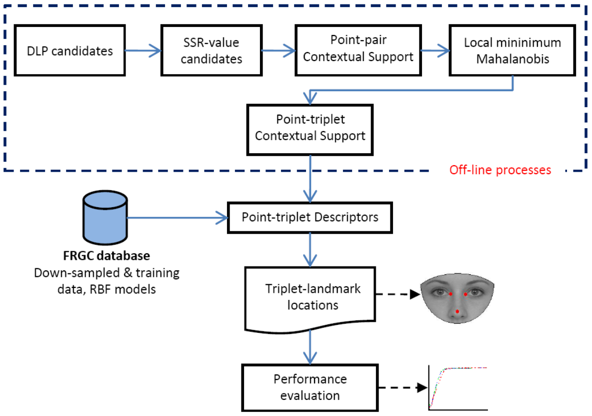

DLP candidates
SSR-value candidates
Point-pair Contextual Support
Local mininimum Mahalanobis
Point-triplet Contextual Support
Off-line processes
FRGC database
Down-sampled & training data, RBF models
Point-triplet Descriptors
Triplet-landmark locations
Performance evaluation

Figure 4. Experimental framework to localise the triplet endocanthions and pronasale landmarks using point–triplet descriptors.

TABLE II. IMPLEMENTATIONS USING POINT-TRIPLET DESCRIPTORS.

	Point-triplet feature descriptor
PT-S1	Weighted-interpolated depth map
PT-S2	Baricenter depth map
PT-S3	7-bins SRS vector
PT-S4	SRS depth map
PT-S5	SRS histogram

TABLE III. THRESHOLDS TO EVALUATE ESTIMATED LOCATIONS.

Success	$error \leq 12mm$
Poor	$12mm < error \leq 20mm$
Fail	$error > 20mm$

Note that localisation is done at the 3D vertex level and we are using a down-sample factor of four on the FRGC dataset [10], which gives a typical distance between vertices of around 3-5mm. Thus, a threshold of 20mm is selected where an estimated landmark would be within a radius of 4 vertices from the ground truth vertex [2].

4. Localisation Performance

Performance figures when using our *point–triplet descriptors* to localise the landmark–triplet pronasale and endocanthions are now presented.

From the block diagram in Figure 4, it can be observed that the point–triplet descriptors localisation performance is related to the candidate triplets obtained off–line. A base–line to estimate the best localisation performance within the point–triplet localisation system is then defined. To compute this base–line, localisation errors between every candidate landmark–triplet are computed against the ground–truth landmark–triplet. For every candidate landmark–triplet their localisation errors are added. Finally, the landmark–triplet with the minimum total localisation error is taken as the best estimation. As observed in Table IV, only the pronasale landmark reaches 100% successful localisation performance, but the same would not be expected for the endocanthion landmarks.

As indicated in Table II, the point–triplet descriptors were embedded into five localisation systems. From these systems, different performance is observed. Hence, a summary of successful localisation is presented in Table IV.

TABLE IV. SUCCESSFUL LANDMARK LOCALISATION SUMMARY.

	Left endocanthion	**Right endocanthion**	**Pronasale**
PT-S1	82.90%	76.03%	98.82%
PT-S2	90.17%	81.92%	99.60%
PT-S3	91.35%	83.98%	99.01%
PT-S4	93.71%	89.98%	99.21%
PT-S5	90.76%	84.67%	99.60%
Base line	96.26%	99.01%	100.00%

5. Discussion and conclusion

This paper devised new surface descriptors, derived from either unstructured surface data, or a radial basis function (RBF) model from the face surface. Then, two new families of descriptors were introduced, generally named as *point–triplet descriptors*, which require three vertices respectively for their computation.

Our *point–triplet descriptors* approach was done based on the belief that a good descriptor must be invariant to pose and orientation. From these criteria, the *7–bins SRS vector* descriptor is the only one that possesses these two properties, making this a potential descriptor for future research.

As for the *point-triplet descriptor* in general, *7–bin SRS vector* and SRS histograms are undirected. *Weighted–interpolated depth map* and *SRS depth map features*, depend on a normal's orientation. Finally, a *Baricenter depth map* feature is undirected, as long as it is binned according to fixed labels from sample points.

This paper presented performance figures when computing every feature descriptor to localise particular facial landmarks as summarised in Table IV. However, this is not the only property that can be observed from them. The motivation to investigate feature descriptors using a number of vertices, e.g. one [2]-[3], two [4], or three, is based on natural limitations associated with each feature descriptor. For instance, a very good question could be: *why use more than one vertex to compute a feature, when SSR histograms or spin–images are able to robustly localise the pronasale landmark?* [1]. There are several reasons that can be discussed in answering this question; however, at this point the focus is on three main arguments:

a) *Robustness to extreme pose variation*

The experimental feature descriptors, computed from a single vertex, e.g. *DLP*, *SSR features* and *spin–images*, are defined radially and a decrease in performance for particular facial landmarks is expected when computed from self occluded data, such as in pure profiles [5]. For instance, an *SSR histogram* at the pronasale landmark in a pure profile will be computed from the half of the nose in the best case, which suggests a reduction in effectiveness. In this respect, *point–triplet descriptors* are flexible and they can be computed from a set of distinctive landmarks, present within a wide range of pose variations, as observed in our preliminary experimentation.

b) *Single facial landmark dependence*

Most 3D face processing applications depend on the pronasale detection to extract the face from a shape image. Although our previous work [5] presents experimental results supporting the pronasale as the most distinctive facial landmark among eleven, those results are from nearly front–pose data, and a different performance is expected using data with pose variations. Thus, the most distinctive facial landmark cannot be depended on alone. Contrarily, with *point–triplet descriptors*, more than one vertex can be combined to assist a landmark localisation.

c) *Scale invariance:*

Computing a feature descriptor based on a single vertex does not provide enough information to define an appropriate scale for an intended facial feature [5]. For instance, SSR histograms are computed from *10 to 45mm* in steps of *5mm* [1]-[5]. Similarly, *distance to local plane* (DLP), *SSR values* and *spin–images* need a specific radius to be computed [1]-[5]. Contrarily, *point–triplet features* are scale invariant, where surface shapes can be encoded within the triangular region defined by the given triplet of points.

However, in exchange for those advantages mentioned above, it is necessary to collect suitable candidates in triplets to compute a *point–triplet descriptor*. This is a crucial task, because the overall system performance greatly depends on these initial candidates.

References

[1] Pears, N., Heseltine, T., and Romero, M. (2010). From 3d point clouds to pose–normalised depth maps. International Journal of Computer Vision, 89:152–176.

[2] Romero, M. and Pears, N. (2008). 3D facial landmark localisation by matching simple descriptors. In IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 08).

[3] Romero, M. and Pears, N. (2009). Landmark localisation in 3D face data. In the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, pages 73–78.

[4] Romero, M. and Pears, N. (2009). Point–pair descriptors for 3D facial landmark localisation. In IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 09).

[5] Romero, M (2010). Facial landmark localisation from 3D face data. PhD Thesis. Department of Computer Science, The University of York, UK.

[6] Ekman, P. (2006). Darwin and facial expressions: a century of research in review. Malor Books.

[7] Pantic, M. and Rothkrantz, L. J. (2000). Automatic analysis of facial expressions: the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1424–1445.

[8] Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. J. Cognitive Neuroscience, 3(1):71–86.

[9] Belhumeur, P., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7):711–720.

[10] Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W (2005). Overview of the face recognition grand challenge. IEEE Int. Conf. on Computer Vision and Pattern Recognition, Volume 1, pages 947–954, Washington, D.C., USA.

[11] Clement Creusot, Nick Pears, Jim Austin (2010). 3D face landmark labelling. Proceedings of the ACM workshop on 3D object retrieval, Italy.

[12] Naoufel Werghi, Haykel Boukadida, Youssef Meguebli (2010). The spiral facets: A unified framework for the analysis and description of 3D facial mesh surface. 3D Research, 1(3), p1-11.

[13] Neşe Alyüz, Berk Gökberk, Lale Akarun (2010). Regional registration for expression resistant 3-D face recognition. IEEE Trans. on Information Forensics and Security, 5(3), p425-440.

[14] Martínez, A. M. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(6):748–763.

[15] Zhao, W. and Chellappa, R. (2005). Face Processing: Advanced Modeling and Methods. Academic Press, Inc., Orlando, FL, USA.